PRACOWNIA INFORMATYCZNA
Lista 2

BIOLOGICZNE BAZY DANYCH

1. Dowiedz się czym charakteryzują się formaty danych:
   a) FASTA
   b) FASTQ
   c) VCF (Variant Call Format)

   Zastanów się jak cechy charakterystyczne tych formatów mogą pomóc Ci w analizie plików, wyodrębnieniu konkretnych informacji itp.

2. Otwórz przeglądarkę internetową i korzystając z bazy danych Ensembl znajdź:

   a) wszystkie sekwencje białkowe dostępne dla człowieka
   b) sekwencje ncRNA dla dingo
   c) adnotacją genomową w formacie GTF dla szczura wędrownego

   Otwórz terminal. Przejdź do katalogu nazwanego dwiema literami Twojego imienia oraz nazwiska (np. Adam Mickiewicz posiada katalog o nazwie „AdMi"). Pobieranie znalezionych danych wykonaj w wierszu poleceń. W terminalu wpisz odpowiednie polecenie wraz z adresem linku prowadzącym do danych. Przyjrzyj się nazwom pobranych plików.

3. Korzystając z plików z poprzedniego zadania:
   a) wyświetl zawartość pliku przechowującego sekwencje niekodujące
   b) rozpakuj plik zawierający sekwencje białkowe bez usuwania oryginalnego pliku.
   c) sprawdź czy znany jest gen o nazwie Pcmt1-201 dla szczura wędrownego. Na którym chromosomie się znajduje?

4. Wiele nowoczesnych programów do mierzenia poziomu ekspresji genów w oparciu o dane RNA-seq potrzebuje jako plik wejściowy wszystkich znanych sekwencji transkryptów dla badanego organizmu. Utwórz katalog o nazwie „transkryptom_hsapiens". Z bazy Ensembl pobierz do niego transkrypty kodujące białko i transkrypty niekodujące u człowieka. Połącz oba pliki w jeden i spakuj go.

5. Plik o nazwie „homo_sapiens_variant_annot.txt" jest wynikiem adnotacji funkcjonalnej 4 polimorfizmów:
   a) jakie to polimorfizmy?
   b) jakie są ich pozycje w genomie?
   c) czy są zlokalizowane w genach?

   Uwaga! Plik jest zapisany w formacie VCF, ale ze względów technicznych ma rozszerzenie „txt".

6. Z pliku o nazwie „bos.txt" wyodrębnij koordynaty polimorfizmów i zapisz je w nowym pliku.